統語解析情報付きコーパス検索用インタフェースの開発

長崎郁† アラステア・バトラー† スティーブン・ライト・ホーン†
プラシャント・パルデシ† 吉本 啓†♠
†国立国語研究所 理論・対照研究領域 ♠東北大学 高度教養教育・学生支援機構
{inaga, horn.s.w, prashant}@ninjal.ac.jp, ajb129@hotmail.com, kei@compling.jp

1 はじめに

国立国語研究所では、現代日本語に関わる、分野を超えた研究や技術開発に役立てるために、統語解析情報をアノテートした初の本格的日本語コーパスである NINJAL Parsed Corpus for Modern Japanese (NPCMJ) を構築している。その一部はすでにウェブ上で公開中であり、併せて、広範囲のユーザーによる使用を目指して、検索用インタフェースも複数公開している。本発表では、これらインタフェースの目的、特色、および意義について解説を行う。

NPCMJ および検索用インタフェースは、以下の国立国語研究所ウェブサイトから公開されている。

http://npcmj.ninjal.ac.jp/

2 NPCMJ とその検索利用

NPCMJ は日本語としては初めての、テクスト中の各文に対して句構造をタグ付けした本格的なコーパスである。自動意味解析を利用することによって、複雑な構文も含めて文中の語句間のすべての依存関係 (dependency) に関する情報を提供できるという特色を持つ (バトラー他 2016)。

一般に、コーパスは検索利用されて初めて意義を発揮するが、そのためにはコーパスが検索に応じることのできる質を保っていなければならないと同時に、それを十分に生かすことを可能にする検索用インタフェースの存在も重要である。特に NPCMJ 利用者の多くは文系の言語研究者であると想定され、言語処理技術を前提としない検索手段の提供が必要である。他方、本コーパスの持つ精緻な文法情報を完全に活用するためには、正規表現を使った統語パターンの検索が不可欠である。そこで本プロジェクトでは、数理的知識を全く必要としないものから統語パターンの検索まで、次のように 5 種類のインタフェースを開発、公開している。

概要とコンテキスト表示: 原テクストとその目録および書誌情報。各文の統語解析木 (ツリー) を表示できる。

パターン・ブラウザー: 事前に準備されたメニュー形式から選ぶことによって、特定の品詞や文型、構文に相当する文を検索し、統語解析木とともに表示する。

文字列検索: 特定の文字列を持つ文を検索し、統語解析木とともに表示する。

クエリ作成: 表の形で示される木構造から必要な情報を選んでクエリを作成することで、パターン検索を支援するシステム。

ツリー検索: カッコ付き木構造、TGrep-lite または XPath のいずれかを入力して行うパターン検索。

3 概要とコンテキスト表示

「概要とコンテキスト表示」では、原テクストをそのまとまりごとに書誌情報とともに提示している。それぞれの文の統語解析木を表示することができる。

なお、すべての種類のインタフェースを通じて、統語解析木の表示は、ディフォールトとしては比較的フラットな木表示として与えられる。この他に、以下の表示モードを選ぶことができる。

インデクス表示 (indexed): 統語解析情報を自動意味解析することにより、指示対象である個体やイベントのアイデンティティ情報をインデクスの形で表示する。これにより、複雑な構文やゼロ代名詞においても、述語の項構造が正確に示される。

表 1: 文字列検索のオプション

	検索表現の最初と最後は単語の切れ目と一致しなくてもよい	検索表現の最初の文字を単語の始まり、最後の文字を単語の終わりと指定
検索表現内部のセグメンテーションを指定しない	Liberal	Character
検索表現内部のセグメンテーションを半角スペースにより指定	Mine	Strict

二分木表示 (binarised): ディフォールトとしての比較的フラットな木構造に代わり、統語解析情報を二分木に変換して表示する。

依存関係表示 (dependency): 主要句 (head) を中心として、他の語句が果たしている意味格役割などの依存関係が示される。

意味表示 (semantics): 統語解析情報から自動的に生成された一階述語論理式を表示する。

意味派生表示 (eval): 一階述語論理式を派生するための意味計算が行われる過程を表示する。

また、統語解析木の表示は単独の例文に限られず、その前後の複数の文のものを同時に表示できるので、文脈を確認しながら木構造を見ることも可能である。

4 パターン・ブラウザー

トップダウン的に事前準備されたメニューの中から選ぶという形式によって、特定の品詞、句カテゴリーおよび構文を持つ文を統語解析木とともに表示する。コーパス利用の初心者にとってはもっとも利用しやすいインタフェースの 1 つであるが、その分個々のユーザーの必要に対し柔軟に対応することはできない。しかし、第 6 節で説明するクエリ作成のための入力として利用することもできる。さらに、パターンの定義に使われている XPath 表現も検索に役立てることができる。

5 文字列検索

検索したい表現のセグメンテーション (分かち書き) が分からない場合や、単語の連鎖に対してどのようなアノテーションが与えられるかを知りたい場合に文字列検索は効力を発揮する。文字列に関するセグメンテーションの様々な可能性に応じて、Liberal, Character, Strict および Mine の 4 種類のオプションが用意されている。表 1 を参照のこと。

また、検索表現の内部に別の文字が挿入されている例を検索できる「よくばり検索」も用意されている。この場合、検索表現は複数の単語へとセグメンテーションが行われていてもよい。挿入される文字数はユーザーにより指定される。

6 クエリ作成

これまでに説明したパターン・ブラウザーにしても文字列検索にしても、一定の有用性はあるが、NPCMJ の精緻なアノテーションを完全に生かすためには抽象的な検索パターンを駆使して検索利用することが必要である。しかし、そのためには形式統語理論等の言語の数理的側面に関する知識が不可欠であり、言語研究者の大部分にとっては敷居が高い。

形式言語理論にも言語処理技術にも習熟しておらず、また当該のツリーバンクのアノテーションの詳細にも通じていないユーザーが統語パターンを使ってツリーバンクを検索することを可能にするウェブ上のユーザーフレンドリーな検索エンジンとして、クエリ作成というインターフェースを提供している。ここでは、上記の 3 つのインターフェースを使って得られた文を元に、クエリが作成できる。クエリ作成は、ICECUP における Fuzzy Tree Fragment 機能 (Nelson, Wallis and Aarts 2002) に倣ったインターフェースで、既成の構造から得られる特徴を自由に選択したクエリの作成を支援する。文を自分自身で作例し、自動解析にかけて、その結果を出発点とすることもできる。これはルーヴァン大学の GrETEL (Augustinus 2016) と類似の機能である。

NPCMJ では、上記の「概要とコンテキスト表示」「パターン・ブラウザー」「文字列検索」のいずれを出発点としても、その検索結果にもとづいてクエリが作

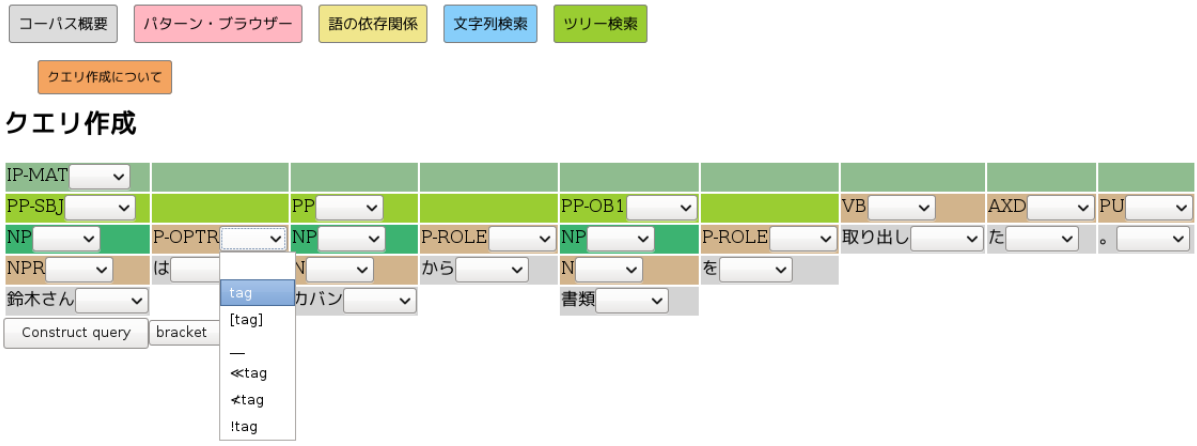

図 1: クエリ作成

成できる。さらに、GrETEL におけるように自然言語の文を自動統語解析し、その解析結果を利用することもできる。

ノード上に付加されたドロップダウンリストの中からオプションを選んでクエリ作成を行う (図 1 を参照のこと)。ドロップダウンリストからは、

(1) 解析結果のタグ全体か、拡張タグ (機能タグ) を除く基部のみか、あるいは拡張タグのみの指定

(2) タグを検索のフォーカスとして指定

(3) 当該のタグが任意の場所に埋め込まれていることの指定

(4) 当該の単語の語形の指定

(5) 否定的条件に関する指定

(6) ワイルドカードとしての指定

が行える。

このようにして一旦表形式のツリーを作成し、ウェブ上で処理すると、クエリの統語解析木表示およびクエリそのものが出力される。クエリの言語は、カッコ付き木構造、XPath、TGrep-lite の 3 つの中から選ぶことができる。この段階でクエリの手修正も可能である。クエリをサブミットすることで、コーパスの用例が検索される。

7 ツリー検索

NPCMJ の本格的利用のためには、統語構造のパターンを入力して行うツリー検索が必要となる。ツリー検索は、テクストボックスにカッコ付き木構造、TGrep-lite、あるいは XPath のいずれかを使ったクエリを入力して行う。

7.1 カッコ付き木構造

ペン・ツリーバンクのカッコ付き木構造の記法 (Santorini 2010) を使って、指定した木構造 (統語解析木) をその一部または全体として持つ文を検索することができる。以下のクエリ

```
(NP (-REL _) (N 人))
```

は、関係節の後に「人」という名詞 (N) が続き、それらが名詞句を構成するような木構造を検索する。ワイルドカード (_) を使って、関係節の内容は何でもよいということを指定している。

カッコ付き木構造検索では、ノードや直接支配関係の否定も指定することができる。しかし、クエリの表現力には限界があり、NPCMJ をより柔軟に利用していくには TGrep-lite や XPath の使用が勧められる。

7.2 TGrep-lite

TGrep-lite は、ツリーバンク中の統語構造を検索するための検索エンジン TGrep (Pito 1993, 1994) にもとづいており、簡単でありながら、オンラインでコーパスを検索するのに適した十分な力を備えている。

TGrep-lite では、ノード間の直接・間接の支配関係、先行/後行関係、姉妹関係、およびそれらの否定によって統語構造木のパターンを指定する。表 2 に TGrep-lite におけるノード間の関係の指定法を掲げる。

表 2: TGrep-lite におけるノード間の関係の指定

A == B	B は A と同一のノードである
A << B	A は B の先祖である
A >> B	A は B の子孫である
A < B	A は B の直接の先祖（親）である
A > B	A は B の直接の子孫（子）である
A .. B	A は B に先行する
A ,, B	A は B に後行する
A . B	A は B の直前に置かれる
A , B	A は B の直後に置かれる
A $ B	A は B の姉妹である
A $.. B	A は B の姉妹であり、かつ、B に先行する
A $,, B	A は B の姉妹であり、かつ、B に後行する
A $. B	A は B の姉妹であり、かつ、B の直前に置かれる
A $, B	A は B の姉妹であり、かつ、B の直後に置かれる

TGrep-lite では、TGrep 言語で書かれたクエリを XPath クエリに書き換えてから XML ツリーを検索することから、統語構造を構成する 3 種類のノード、すなわち、終端ノード (単語)、品詞ノード、および句ノードを区別する必要がある。ノードラベルの選択肢を表す時は、これら異なるタイプのラベルを混ぜてはならない。また、選言の使用も限定される。

クエリの文字列が完全に品詞タグと一致する場合、品詞ノードを指定していると優先的に解釈される。それ以外の場合、文字列が句レベルのタグの基部または拡張部と一致していれば、句ノードを指定していると解釈される。それ以外の場合、文字列が品詞タグの拡張部と一致していれば、品詞ノードを指定していると解釈される。これらのいずれでもない場合、文字は語形/見出し形の全体または一部を指定していると解釈される。

7.3 XPath

Alpino XML 形式 (van Noord 2013) のコーパス・データに対して、XML ドキュメント中の特定のノードへのアクセスを可能にする記述言語である XPath を利用して検索を行うことも可能である。

XPath による検索は、まず特定の軸 (axis) ―先祖、子孫、兄弟、自分自身など―を指定してノードを選択した上で、そのノードの持つ cat 属性 (文法カテゴリーを表す)、pt 属性 (品詞を表す)、begin 属性および end 属性 (語順を表す) を指定することで行う。begin/end 属性は Alpino XML の大きな特色であり、これによりノードの線形順序を記述することができる。これを利用することで、語順や隣接関係についての検索が可能になるのである。

8 おわりに

NPCMJ とともに開発、公開されているインタフェースについて解説を行った。

これらのインターフェースは、コーパス全体のあり方やアノテーション方式の適用について知るには役に立つのは事実であるが、特定の情報の抽出や込み入った条件の検索の場合、コーパス全体或いは一部をダウンロードし、off-line ツールで処理するほうが適切な場合が多いことを付記しておく。

参考文献

Augustinus, L. (2016) About GrETEL. http://Gretel.ccl.kuleuven.be/project/

アラステア・バトラー・吉本啓・岸本秀樹・プラシャント・パルデシ (2016)「統語・意味解析情報付き日本語コーパスのアノテーション」『言語処理学会第 22 回年次大会発表論文集』, pp. 589-592.

Nelson, G., Wallis, S. and Aarts, B. (2002) *Exploring Natural Language: Working with the British Component of the International Corpus of English.* John Benjamins.

Pito, R. (1993, 1994) *Tgrep (1).*

Santorini, B. (2010) Annotation Manual for the Penn Historical Corpora and the PCEEC (Release 2). Tech. rep., Dep. of Computer and Information Science, University of Pennsylvania.

van Noord, G. et al. (2013). Large Scale Syntactic Annotation of Written Dutch: Lassy. In: P. Spyns and J. Odijk (eds.) *Essential Speech and Language Technology for Dutch: Resources, Tools and Applications.* Springer.